Introduction:

- Photos are images that represent a single captured instance in time
- Stringing photos together can create the framework of a story
- A story is not a sequence of captions (Figure 2, top row); it requires natural, abstract language and, more importantly, temporal context
- There is a difference between an isolated caption and one that is written as a part of a story sequence; this is the difference between, say, “a rock hovers above water” and “a stone was skipped across the lake” (Figure 1)
- In order to tell a story, the captions should follow natural language conventions and factor each photo’s order in the sequence of photos
- Computers have been able to generate isolated captions for photos [1][2], but currently struggle to create a sequence of captions that tell a cohesive story

Dataset:

- Dataset consists of a photo training dataset (includes over 80,000 unique photos in over 20,000 sequences) and accompanying information on the photos, which comes in the form of descriptions-in-isolation (DII), descriptions-in-sequence (DIS), and stories-in-sequence (SIS) [3]
- DII, DIS, and SIS contain crowdsourced captions (annotations) for photos in the training set [4]
- Multiple captions can exist for a single photo, enriches wealth of natural language vocabulary and interpretation of a photo
- Tiers show the effect of temporal context on each photo (Figure 2)
- Multiple captions (annotations) can exist for a single photo, demonstrating the different ways people can interpret an image
- DII and SIS annotations are available from the VIST site, but DIS is not [3]

Goal:

- For a given sequence of photos, generate a story by creating captions using natural language and factoring in the passage of time (Figure 3)
- Currently working with photos provided in the Visual Storytelling (VIST) dataset, but would like to be able to use new photos from beyond VIST as input for this program

Method:

- Build an artificial intelligence (AI) that uses a recurrent neural network (RNN) to train the story generation program on the training dataset [4]
- Run the RNN on a fraction of the dataset, then hopefully be able to run it on the full dataset
- Caption input photos by running them through the RNN; categorize them based on the VIST training set, then generate captions based on annotations associated with the categorization of the photo in the SIS portion of the dataset
- Caption generation uses the techniques used in the Vinyals et al. [1] and Devlin et al. [2] for single image caption generation

Going Forward:

- Recreate the Vinyals et al. [1] and Devlin et al. [2] caption generators; build a convolutional neural network (CNN) that classifies the input image; build a long short-term memory (LSTM) RNN that takes into account the image input and previous word in caption so far
- Run the output of the above LSTM RNN through the proposed RNN in VIST [4]

References